Calibration of the 2UP model

Andree, B.P.J.; Koomen, E.

2017

Calibration of the 2UP model

Bo Pieter Johannes Andrée
Eric Koomen

December 22, 2017

COLOPHON

TITLE

Calibration of the 2UP model
Spinlab Research Memorandum SL-13

AUTHORS

Bo Pieter Johannes Andrée, Spatial Information Laboratory (SPINlab), Vrije Universiteit Amsterdam.
Eric Koomen, Spatial Information Laboratory (SPINlab), Vrije Universiteit Amsterdam.

CONTACT

Vrije Universiteit Amsterdam
Faculty of Economics and Business Administration
Department of Spatial Economics/ Spatial Information Laboratory (SPINlab)
De Boelelaan 1105
1081 HV Amsterdam
Netherlands
Phone: +31 20 5986095
Email: e.koomen@vu.nl
Website: https://spinlab.vu.nl/

This report was commissioned by PBL Netherlands Environmental Assessment Agency, Den Haag.

Contents

1. Introduction

The 2UP model – to an Urban Preview – is developed for the spatially explicit simulation of the future growth of cities and population at a global scale. The model describes urban land use at a fine 30" spatial resolution (approx. 1km near the equator). In the current version of the model, the allocation of urban land use is mainly based on current population densities. To improve this calibration PBL has requested Vrije Universiteit Amsterdam to apply a more advanced, statistics-based calibration procedure that links observed land-use patterns to spatially explicit drivers of urban development. The output of this calibration effort can be used to define the suitability for urban land use in the 2UP model. This short memo describes the method applied in calibration and discusses the results.

1.1 AutoGLM calibration

The calibration of the 2UP model was performed using the R package 'autoGLM' that was developed by Bo Andrée for automated Generalized Linear Models suitable for large datasets. The package resulted from the needs for automated procedures for the statistical calibration of the LUISA model framework that is used by the Joint Research Centre of the European Commission for ex-ante evaluation of policies. The code is open-source, and can be installed on any machine by typing into your R terminal:

```
library(devtools)
install_github("BPJandree/AutoGLM")
```

AutoGLM performs minimization of Information Criteria to produce an optimal model containing only powerful predictors. The code can handle large datasets by using smart sampling strategies that ensure that subsets of the data are representative of the entire dataset. In this way it is able to train models without the need of running on the entire dataset. The package is also useful in cases where a large number of variables is available, and the researcher does not want to impose a-priori which variables should be kept out of a model. The current package allows for linear, probit and logit models. In this calibration we work with the logit model. Logit models are well-documented in the literature including many basic text books on econometrics, we refrain from technical discussions as we assume the reader is familiar with the logistic regression framework (as, for example, discussed in an earlier PBL-report: Loonen and Koomen, 2009).

The major challenge in statistical calibration of the global 2UP model is the size of the data. After dropping NoData values, there are still over 300 million observations, this means that the model needs to process around over 2 billion data values (considering the 7 variables included in the analysis). Cloud processing solutions are available for such large operations, but these are often costly. Moreover, it is important to consider that not all data is needed to produce a global statistical model. Therefore, we use the sampling strategy from the autoGLM package to draw samples with near identical statistical properties of the entire data set. This strategy repeatedly draws random samples and performs t-tests and F-tests to compare means and variances between the sample and the population data. In this way we discard 75% of the data. After this we are still left with approximately 80 million observations, which is still a fairly large dataset.

We use the default autoGLM approach to perform a sampling round and train the model on 25% of the preselected data, and compare in-sample and out-of-sample predictions. The model is thus trained on approximately 20 million observations, and the prediction results are compared against approximately 60 million out-of-sample observations. Due to the size of the vectors and matrices involved, significant RAM is required. The model fitting was performed on a 64GB RAM machine, at the peak over three quarters of the RAM was utilized.

1.2 Explanatory variables

Spatially explicit data sets capturing important drivers for urban development (independent variables) were provided by PBL. The general characteristics of the data are listed in Table 1. Before starting the analysis, the correlation between the provided variables was tested. The resulting correlation table indicates that the independent variables are not strongly correlated (Table 2).

Table 1 Descriptive statistics of the variables included in the statistical analysis (initial PBL file names between brackets)

Variable	**Min**	**Max**	**Mean**	**St. dev.**
Urban area 2010 (dependent variable)	0	1	0.002	0.045
Distance to urban area 2010 (index value, abbreviated to UrbanPot)	0	1	0.002	0.030
Elevation (in metres)	-406	8519	1139.6	1158.3
Slope (in degrees)	0	51.6	1.3	2.4
Terrain roughness index (TRI)	0	80	30.5	34.7
Travel time	0	11.8	0.137	0.776
Distance to coast (UrbanCoastPot)	0	29.7	0.004	0.159

Note: NoData values were dropped, total number of remaining observations = 77,392,119

Table 2 Correlation matrix of the included variables

	Urban area	**Dist. urban**	**Elevation**	**Slope**	**Terrain r.i.**	**Travel time**	**Dist. coast**
Urban Area		0.781	-0.035	-0.012	0.048	0.355	0.359
Dist. to urban area			-0.051	-0.015	0.065	0.496	0.422
Elevation				0.071	-0.506	-0.119	-0.024
Slope					-0.330	-0.002	-0.006
Terrain roughness ind.						0.120	0.032
Travel time							0.207
Dist. to coast							

2. Results

Overall, we find that for the current global dataset, all provided variables contribute significantly to the predictive power of the model (Table 3). The performance of this relatively simple model is fairly impressive. The overall accuracy is not so interesting as it is strongly influenced by all the correctly predicted non-urban sites (and most of the area is not urban). More interesting are the scores for the correctly predicted urban sites: 63.8% of the observed urban sites are correctly predicted by the model (this metric is often referred to producer's accuracy, see Story and Congalton, 1986), while 80.7% of our predicted sites are actually urban (aka user's accuracy). This implies that we somewhat underestimate the urban area in our predictions. The differences between the within sample results (based on the data used for training the model) and out of sample results (based on remaining observations) are negligible, indicating that the sampling procedure did not introduce additional inaccuracy (Table 4).

Table 3 Results regression analysis

Variables	Estimate	Std. Error	z-value	P value
Intercept	-8.25089	0.03411	-241.89	<2e-16
Distance to urban area 2010	16.62688	0.06297	264.06	<2e-16
Elevation	-0.00066	0.00002	-29.48	<2e-16
Slope	0.07033	0.00594	11.84	<2e-16
Terrain roughness index	0.01376	0.00041	33.34	<2e-16
Travel time	0.12743	0.00284	44.85	<2e-16
Distance to coast	0.12631	0.00852	14.83	<2e-16

Note: all estimates are significant at the 0.1% level (P< 0.001)

Table 4 Predictive power of the model

	Overall accuracy	At observed urban sites	At predicted urban sites
Within Sample	0.999	0.638	0.807
Out of Sample	0.999	0.639	0.806

Note: This assessment is based on applying a 50% probability threshold for defining the cells that are expected to become urban. Results will change when cells with a lower probability threshold is chosen.

The most important driving force in the current analysis is distance to the urban area in 2010. This variable is not a straightforward distance calculation, but captures the amount of urban area surrounding a grid cell in a similar way as focal statistics do in ArcGIS. In this case the indicator sums up the total amount of urban area in the neighbouring grid cells after applying a relative weight based on their distance to the central grid cell. The eight directly neighbouring grid cells have a total weight of around 33% and this weight quickly decreases with increasing distance (Figure 1). Elevation has a considerable impact at higher altitudes: at 1000 metres the probability of a cell being urban is around 60% lower than at sea level. Slope has a slightly counter-intuitive impact as the results suggest that steeper slopes are more attractive for urban development. This may relate to the fact that slopes are generally mild at the 1km resolution of the data (mean slope is

1.2 degrees with a standard deviation of only 2.4 degrees). Steep slopes are extremely rare and will most likely coincide with high elevations were urban development is unlikely anyhow. For the remaining variables we lack information on their exact contents to interpret the results. During the final discussion meeting we are happy discuss this part of the results in more detail.

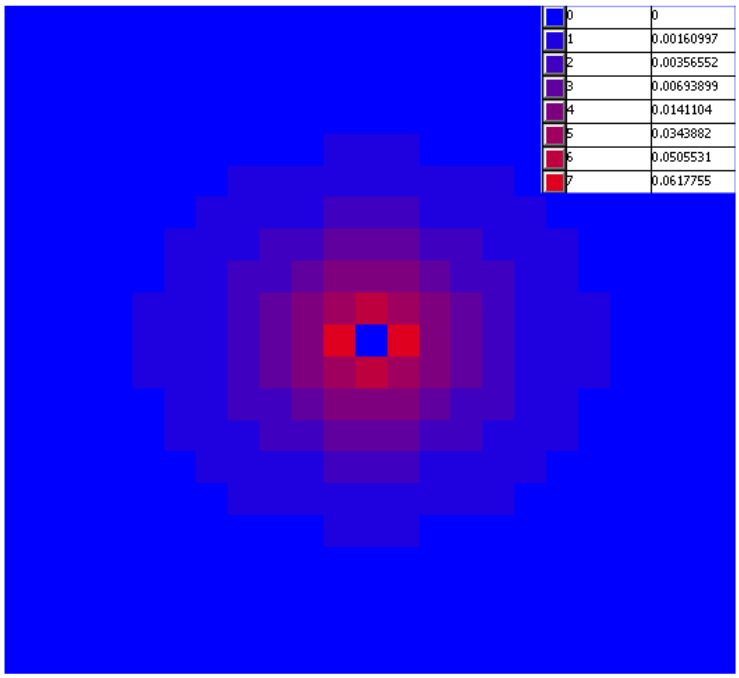

Figure 1 The relative weight of individual grid cells surruonding a central grid cell as applied in the distance to urban area indicator

3. Discussion and possible next steps

3.1 Application of results in the 2UP model

The coefficients of the regression analysis can directly be used to define the suitability maps needed for land-use simulation. In the classic (continuous or logit) version of Land Use Scanner allocation follows this basic equation (for more details, see: Koomen et al., 2011):

$$M_{cj} = a_j * b_c * e^{S_{cj}}$$

in which:

- M_{cj} is the amount of land in cell c expected to be used for land-use type j;
- a_j is the demand balancing factor that ensures that the total amount of allocated land for land-use type j equals the sector-specific claim;
- b_c is the supply balancing factor that makes sure the total amount of allocated land in cell c does not exceed the amount of land that is available for that particular cell;
- S_{cj} is the suitability of cell c for land-use type j, based on its physical properties, operative policies and neighbourhood relations.

If the 2UP model follows this approach, the suitability map for urban land use can be created by simply adding up the values for the intercept and the local values for all included variables multiplied by their estimated coefficients.

3.2 Potential improvements

The set of explanatory variables provided by PBL will most likely be helpful in generating plausible simulation results and we are willing to assist in transferring the obtained regression coefficients into suitability maps and help interpret and document results. To conclude this report we share some thoughts on the calibration process and possible improvements.

The autoGLM package allowed estimating a basic logistic regression model for the very large dataset required for the calibration of the 2UP model. The autoGLM sampling procedure worked fine in this case. For future applications it is worth considering adapting the sampling strategy to ensure that non-urban cells are sampled proportionally from different types of land use or environments. In this analysis they were picked from the preselected data set that only described land use as being urban or not. By including reference to, for example, other land-use data sets (e.g. MODIS or CCI-LC, see Diogo and Koomen, 2016) it is possible to create a sampling strategy that sampled proportionally from different land-use types. Considering the large number of observations that was used for this calibration we do not expect this issue to influence the presented results.

Should larger sets of explanatory variables or more advanced regression techniques (e.g. interaction variables, spatial lags) be considered for calibration, having sufficient RAM for computation will become an issue. Possible solutions in this case are the application of cloud

computing or the estimation of several region-specific models that require on lower numbers of observations. The R code of the autoGLM package can be adjusted to incorporate such solutions or implement interaction variables and spatial lag specifications.

The current statistical analysis assesses the importance of a small set of drivers for explaining current urban patterns. This assessment can be used to build suitability maps that are able to replicate current urban development patterns and – when an additional demand for urban area is added to the model –also to simulate future development patterns. It is important to consider, however, that current urban land use is not likely to disappear and future developments are likely to be additions to currently existing urban areas. An alternative option would be to define an urban development model that focusses on simulating changes in urban area rather than replicating the total urban area of the future. To underpin such models it is essential to analyse urban change rather than current, static urban patterns.

In this analysis distance to urban area seems to be the most important variable. This variable will help reproduce larger urban areas (of several squared kilometres) based on their spatial clustering, but will be of less value to replicate more isolated smaller urban areas consisting of a few individual grid cells as these lack urban grid cells in the vicinity to indicate the location being suitable. The problem of disappearing isolated urban features can, of course, be controlled by setting a resistance against change (inertia) for existing urban area.

To further improve the statistical calibration of the model we suggest to differentiate between different regions (continents) as we expect the drivers in, say, good old Europe to differ from those in more dynamic regions of the world such as Asia or Africa. Specifying the suitability maps at the level of more coherent regions could allow to address differences in the strengths of particular drivers reflecting differences in location preferences, planning traditions, commuting behaviour etc. Ideally such more refined analyses should be complemented with the inclusion of more detailed and region-specific data sources related to, for example, accessibility, economic opportunities, soil types, climatic conditions (see, for example, Motamed et al., 2014; Seto et al., 2011).

References

Diogo, V., Koomen, E., 2016. Land Cover and Land Use Indicators; review of available data. OECD Publishing.

Koomen, E., Hilferink, M., Borsboom-van Beurden, J., 2011. Introducing Land Use Scanner, in: Koomen, E., Borsboom-van Beurden, J. (Eds.), Land-use modeling in planning practice. Springer, Dordrecht, pp. 3-21.

Loonen, W., Koomen, E., 2009. Calibration and validation of the Land Use Scanner allocation algorithms, Bilthoven.

Motamed, M.J., Florax, R.J.G.M., Masters, W.A., 2014. Agriculture, transportation and the timing of urbanization: Global analysis at the grid cell level. Journal of Economic Growth 19, 339-368.

Seto, K.C., Fragkias, M., Güneralp, B., Reilly, M.K., 2011. A Meta-Analysis of Global Urban Land Expansion. PLOS ONE 6, e23777.

Story, M., Congalton, R.G., 1986. Accuracy assessment: a user's perspective. Photogrammetric Engineering and remote sensing 52, 397-399.